DOCUMENT RESUME

ED 353 340 TM 019 406

AUTHOR Jarrell, Michele Glankler
TITLE A Comparison of Two Procedures, the Mahalanobis Distance and the Andrews-Pregibon Statistic, for Identifying Multivariate Outliers.
PUB DATE Nov 92
NOTE 27p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Knoxville, TN, November 11-13, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; Computer Simulation; Error of Measurement; Mathematical Models; *Multivariate Analysis; Research Problems
IDENTIFIERS *Andrews Pregibon Statistic; *Mahalanobis Distance Function; Outliers; Repeated Measures Design; Standard Deviation

ABSTRACT

This repeated measures factorial design study compared the results of two procedures for identifying multivariate outliers under varying conditions, the Mahalanobis distance and the Andrews-Pregibon statistic. Results were analyzed for the total number of outliers identified and number of false outliers identified. Simulated data were limited to three dimensions. Using known population parameters, 1,100 samples of size 150 were computer-generated. Outliers were induced by replacing randomly selected data points in each sample with plus or minus the value of three or six standard deviations (SDs) from the mean. The samples had outliers induced into one dimension, then into two dimensions, and finally, into three dimensions. Each sample was manipulated in terms of the number of dimensions and the extremeness of the outliers. Six outlier samples were produced from each original sample. Analyses of variance were run using the number of false outliers and then the number of total outliers as the dependent variable. Tukey's honestly significant difference was calculated for each of the main effects. Each procedure accounted for less than one percent of the total variability in the model. The degree of outlier extremeness accounted for most of the variability in each case, with the three SD group accounting for more outliers than the six SD group. Other than calculation ease, there is no reason to choose one procedure over the other. Included are 8 tables and 39 references. (RLC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED353340

A COMPARISON OF TWO PROCEDURES, THE MAHALANOBIS DISTANCE AND THE ANDREWS-PREGIBON STATISTIC, FOR IDENTIFYING MULTIVARIATE OUTLIERS

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

MICHELE G. JARRELL

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

by

Michele Glankler Jarrell
Evaluation and Assessment Laboratory
The University of Alabama

Paper presented at the annual meeting of the Mid-South Educational Research Association, Knoxville, Tennessee, November 11 - 13, 1992

TM019406

BEST COPY AVAILABLE

A COMPARISON OF TWO PROCEDURES, THE MAHALANOBIS DISTANCE AND THE ANDREWS-PREGIBON STATISTIC, FOR IDENTIFYING MULTIVARIATE OUTLIERS

INTRODUCTION

Examination of outliers is an essential part of any analysis, univariate or multivariate. The results of many classical statistical procedures can be distorted by the occurrence of outliers; "estimators that are optimal under a Gaussian [normal] assumption are very vulnerable to the effects of outliers" (Wainer, 1976, p. 285); therefore the identification and possible removal or accommodation of outliers are important considerations. Often the outliers may be the point of interest, as in identifying exceptional schools in a system or exceptional teachers in a field. An outlier may simply be the result of an error in observation or data entry; in this case, identification would permit the researcher to make appropriate corrections. Gnanadesikan (1977) stated that "the consequences of having defective responses are intrinsically more complex in a multivariate sample" (p. 271) than in a univariate sample.

"Outliers occur very frequently in real data" according to Rousseeuw and Leroy (1987, p. vii). Huber (1977) stated that having five to ten percent "wrong values" (p. 3) is the norm. Thus, procedures for finding these outlying values and for deciding how to handle them are essential. Whatever the cause of the outliers, we must identify them in order to decide how best to deal with them in the statistical analysis. Wood (1983) saw three possibilities for dealing with outliers after they have been identified: remove them from the data set, keep them in order to extend our range of knowledge, or modify the model to accommodate them. Chatterjee and Hadi (1988) recommended the outliers be examined for "accuracy . . ., relevancy . . ., or special significance" (p. 182).

Outliers require consideration; as several authors have indicated, they are an unavoidable problem (Barnett & Lewis 1978; Douzenis & Rakow, 1987; Huber, 1977). If one deals with data, one must be able to identify outliers and to decide how to treat them. Data are often scanned into a computer file for analysis thereby becoming "invisible" (Gentleman & Wilk, 1975, p. 387), or data

are in such large quantities that they are impossible to inspect visually. Procedures to identify possible outliers in large multivariate data sets are a necessity.

Definitions of Outlier

Many of the researchers who have dealt with the problem of outliers have based their work on a subjective definition of an outlier. Outliers have been seen as values that are "dubious in the eyes of the analyst" (Dixon, 1950, p. 488) or that appear "to deviate markedly from other members of the sample" (Grubbs, 1969, p. 1); and Elashoff and Elashoff (1970) stated outliers are observations that are "extreme in some sense" (p. 4). Many other researchers have used the same basic definition (Barnett, 1978; Barnett & Lewis, 1978; Pascale & Lovas, 1976; Rasmussen, 1988; Robertson, 1987).

Guttman (1973) saw an outlier as a spurious observation that did not come from a $N(\mu,\sigma^2)$ population. Gentleman and Wilk (1975) pointed out that an outlier could be an outlier only "relative to some prespecified model or theory . . ." (p. 389). Hawkins (1980) defined multivariate outliers as "values with high probabilities of occurring where the probability density of the true distribution is low, remote from the main body of data" (p. 104).

Many definitions are based on the type of analysis; for instance, in a regression analysis, an outlier is a value which deviates from the regression line (Douzenis & Rakow, 1987) or one with a high residual (Chatterjee & Hadi, 1988). Rousseeuw and van Zomeren (1990) see outliers as "observations that deviate from the model suggested by the majority of the point cloud "(p. 651).

Causes of Outliers

Among the most commonly cited reasons for the occurrence of outliers in data are errors in collecting, recording, coding, or entering data, and deviations from the experimental design (Chatterjee & Hadi, 1988; Douzenis & Rakow, 1987; Portnoy, 1988; Seber, 1984); Barnett and Lewis (1978) referred to these as human error and ignorance. These are the outliers that require identification in order to be corrected or rejected. Some outliers occur due to violations of the

assumptions; they may indicate the model is not an appropriate one for the data, and they will affect the inferences drawn from the procedures used. Outliers may be due to the "variability inherent in the data" (Grubbs, 1969, p. 1) as with data from a "heavy tailed distribution such as Student's *t*" (Hawkins, 1980, p. 1); in this case, the "outliers" are actually valid data points and should not be deleted. Data may actually be from two populations with different distributions, in which case the outliers would be observations not from the basic distribution. These outliers should be rejected or given small weights (Hawkins, 1980).

Identification of Outliers

Identification of outliers is critical because "many of the standard multivariate methods are derived under the assumption of normality and the presence of outliers will strongly affect inferences made from normal-based procedures" (Schwager & Margolin, 1982, p. 943).

There are several procedures for the identification of multivariate outliers, each with its adherents and detractors (Comrey, 1985; Grubbs, 1969). The question addressed in this study will be how two procedures, the Mahalanobis distance and the Andrews-Pregibon statistic, compare in detecting outliers in a multivariate normal population under varying conditions.

Purpose of the Study

Presently many authors suggest that researchers use one or more procedures for identifying outliers before performing their statistical analyses (Gnanadesikan & Kettenring, 1972; Krzanowski, 1988; Stevens, 1984). Outliers and influential data can be the most important data in an analysis and are deserving of special attention according to Gray (1989). The points classified as outliers may differ dramatically according to which identification procedure is used.

This study used a repeated measures factorial design to compare the results of two procedures for identifying multivariate outliers under varying conditions. Results were analyzed for the total number of outliers identified and for the number of false outliers identified. Simulated data were

generated by computer and were limited to three dimensions. Using known population parameters, 1,100 samples of size 150 were generated. Outliers were induced by replacing randomly selected data points in each sample with plus or minus the value of three or six standard deviations from the mean. The samples had outliers induced into one dimension; then the samples had outliers induced into two dimensions; and finally the samples had outliers induced into all three dimensions. Both procedures were used on all samples.

Definition of Terms

Outlier. An extreme observation, either high or low, which does not conform to the distribution of the majority of observations in the sample. An observation from a distribution with different parameters than the majority of the observations. For the purpose of this study, an outlier will be defined as a data point plus or minus either three or six standard deviations.

False Outlier. An observation occurring naturally in the population but identified by the procedure as an outlier.

Hypotheses

The study tested the following null hypotheses using as dependent variables both the total number of outliers detected and the number of false outliers detected. The total number of outliers included both the induced outliers and the false outliers.

H_{O_1}: There will be no significant differences at the .05 level between the two procedures in detecting outliers.

H_{O_2}: There will be no significant differences at the .05 level between the "Three Standard Deviation" and the "Six Standard Deviation" outlier groups in the number of outliers detected.

H_{O_3}: There will be no significant differences in the number of outliers detected at the .05 level among the groups with outliers in one, two, and three dimensions.

Method

Type of Study

This was an empirical study that used computer generated data. Shapiro, Wilk, and Chen (1968) found that "empirical sampling using a high speed computer can provide a very useful general guide on sensitivity properties even with a few Monte Carlo runs (e.g., 100, 200, or 500)" (p. 1371). Two methods of detecting multivariate outliers, the Mahalanobis distance and the Andrews-Pregibon statistic, were compared under varying conditions for a set of population parameters. The population was a large three dimensional multivariate normal data set with known means and variances.

Generation of Data

The data set was generated using a FORTRAN program on the IBM 3090/400E computer at The University of Alabama Seebeck Computer Center. Morris (1975) recommended using his FORTRAN program to generate data for use in Monte Carlo studies; the program creates a population with the desired centroid and covariance matrix. Using the population parameters, 1,100 random samples were generated. Using 1,100 samples gave a bound on the error of between 3 and 3.1 percentage points. The 1,100 random samples from the population were generated on the computer after setting the population parameters in the FORTRAN program. In order to generate data that were consistent with the type of data a behavioral scientist would see, the population parameters used in the study were those of the standard score distribution, that is, multivariate normal (0,1). These population parameters provided for simplicity of interpretation.

The sample sizes were 150. In order to induce outliers in the samples, the following method was used. The standard deviations obtained from the selected population parameters were multiplied by three and by six to obtain constants that replaced, as a positive or negative value, eight

observations in the original samples. Since the samples were randomly generated, the first eight observations in each sample were replaced. For the samples with outliers in one dimension, the appropriate constant (3σ or 6σ of the first variable) replaced the first variable of four observations, and the negative of the appropriate constant replaced the first variable of four observations. For the samples with outliers in two dimensions, the appropriate constants (3σ or 6σ of the first and second variables) replaced the first two variables of two observations; the negative of the constant replaced the first two variables of two observations; the constant replaced the first, and the negative constant replaced the second variable of two observations; and the negative constant replaced the first, and the constant replaced the second variable of two observations. For samples with outliers in three dimensions, each of eight multivariate outliers was induced according to a different pattern. The appropriate constants (3σ or 6σ of the three variables) replaced all three variables in one observation, and the negative constants replaced all three variables in one observation. In one observation, the constants replaced the first two variables, and the negative constant replaced the third; in one observation, the constants replaced the first variable, and the negative constant replaced the last two; in one observation, the negative constant replaced the first variable, and the constants replaced the last two; in one observation, the negative constants replaced the first two variables, and the constant replaced the third. In one observation, the constants replaced the first and third variables, and the negative constants replaced the second; in one observation, the negative constants replaced the first and third variables, and the constant replaced the second. Table 1 illustrates the pattern for inducing outliers.

Each sample was manipulated in terms of the number of dimensions and the extremeness of the outliers; therefore, there were actually six outlier samples produced from each original sample.

Table 1

Clustering of Outliers According to Dimensions

Dimension		
One	Two	Three
+	+ +	+ + +
+	+ +	− − −
+	− −	+ + −
+	− −	− − +
−	+ −	+ − −
−	+ −	− + +
−	− +	+ − +
−	− +	− + −

where + is $x_{(i)} = +3\sigma$ or $x_{(i)} = +6\sigma$ and
where − is $x_{(i)} = -3\sigma$ or $x_{(i)} = -6\sigma$

The data were generated on an IBM 3090\400E mainframe computer using a REXX executable file (International Business Machines, 1988a; International Business Machines, 1988b) to run several SAS programs (SAS Institute Inc., 1990a; SAS Institute Inc., 1990b) and a FORTRAN program. Eleven hundred samples of $n = 150$ and dimension = 3 were generated from a multivariate normal (0,1) population using a FORTRAN program developed by Morris (1975). The Mahalanobis distance (Stevens, 1984) and the Andrews-Pregibon statistic (Andrews & Pregibon, 1978) were calculated for each observation in each sample. A data file was created with the statistics and the row number from the matrix for each of the two procedures for each sample.

For each sample the REXX executable file wrote a one-line file containing an eight-digit seed number. The seed number was read by a SAS program that generated a random number and wrote out a data file that set the parameters for the FORTRAN program. These parameters included the random seed number, the number of decimal places requested, the desired variance-covariance matrix, the vector of desired means, the number of variables and the number of observations, and the

designated output file names. An executable file ran the FORTRAN program that generated a listing of the data; this file was read by a SAS program that transformed the data into SAS/IML matrix form and wrote the matrix in a data file. The SAS program that calculated the Mahalanobis distance was edited using an executable file that substituted the new data matrix during each run through the programs. When the SAS program was run, it induced the outliers into the data matrix, calculated the Mahalanobis distances, and produced a listing of the distances for the matrix. Another SAS program read the output and appended the data to a data file if the Mahalanobis distance met or exceeded the critical value; therefore, after the 1,100 runs through the REXX executable file, there was one data file containing the Mahalanobis distance and row number for each observation identified as an outlier. The SAS program that ran the Andrews-Pregibon statistic on the matrix was edited using an executable file that substituted the new data matrix during each run through the program. When the SAS program was run, it deleted one observation at a time from the data matrix, calculated the Andrews-Pregibon statistic, and produced a listing of the statistics and the row number from the matrix. Another SAS program read the output and appended the data to a data file if the Andrews-Pregibon statistic was equal to or less than the critical value; therefore, after the 1,100 runs through the REXX executable file, there was one file containing the Andrews-Pregibon ratio and the row number for each observation identified as an outlier.

The two data files containing the Mahalanobis and Andrews-Pregibon outlier data were downloaded to a microcomputer. Two BASIC programs were run. The first program read the Mahalanobis outlier data and produced another data file containing the sample number and the number of false outliers, induced outliers, and total outliers identified under each combination of extremeness and dimensionality. Thus, for each sample there were six lines in the output data file. The second BASIC program did the same thing for the Andrews-Pregibon outlier data. Data lines in the output files were coded "1" for Mahalanobis data and "2" for Andrews-Pregibon data, then the

Andrews-Pregibon data were appended to the Mahalanobis data forming a single data file. This file was uploaded to the mainframe and sorted by procedure and sample. The sorted file contained twelve lines of data for each of the 1,100 samples.

Outlier Identification Procedures

The first procedure, the Mahalanobis distance, gives the "distance from the case to the centroid of all cases for the predictor variables" (Stevens, 1984, p. 339). A large distance in relation to the other distances obtained indicates an outlier. This distance also measures an observation's leverage (Chatterjee & Hadi, 1988). The Mahalanobis distance, D_i^2, is expressed in terms of the covariance matrix S:

$$D_i^2 = (X_i - \bar{X})^T S^{-1} (X_i - \bar{X})$$

In this formula, X_i is the vector of data for case i, and $\bar{X}$ is the vector of means for the predictors. Under the assumption that the predictors came from a multivariate normal population, the critical values of D_i are given by Barnett and Lewis (1978) for p = 2 to 5 and for alpha = .05 and .01.

For each sample, the induced outliers were added in one dimension for the first degree of extremeness (i.e., three standard deviations), the Mahalanobis distance was calculated, and a vector of the values was printed. The vector of Mahalanobis distances was read, and any value equal to or exceeding the critical value of 7.81473 (Tabachnick & Fidell, 1983, p. 479) was printed with the row number indicating which observation it had identified. The same procedure was followed for each of the dimensions and each degree of extremeness. Therefore, for each of the 1,100 samples, there was a listing of the possible outliers with their corresponding Mahalanobis distance for each of the six possible combinations of dimensionality and extremeness.

The second procedure, the Andrews-Pregibon statistic, is based on the volume of the confidence ellipsoid. Andrews and Pregibon (1978) stated that this procedure identifies observations

that are potential outliers and that are influential on the linear model estimates. The Andrews-Pregibon ratio is expressed as:

$$AP = \frac{\det(X_{(i)}^{T} X_{(i)})}{\det(X^{T}X)}$$

where *det* is the determinant of the matrix that results from multiplying the two matrices, X^T is the transpose of the X matrix, $X^T_{(i)}$ is the transpose of the X matrix with the ith observation deleted, and $X_{(i)}$ is the X matrix with the ith observation deleted.

For each sample the induced outliers were added in one dimension for the first degree of extremeness, i.e., three standard deviations, the Andrews-Pregibon statistic was calculated, and a vector of the values was printed. The vector of Andrews-Pregibon statistics was read, and any value less than or equal to the critical value of 0.9484 (Jarrell, 1991) was printed with the row number indicating which observation it had identified. The same procedure was followed for each of the dimensions and each degree of extremeness. Therefore, for each of the 1,100 samples, there was a listing of the possible outliers with their corresponding Andrews-Pregibon statistic for each of the six possible combinations of dimensionality and extremeness.

Data Analysis

Repeated measures analysis of variance was run using the total number of outliers identified in each sample by the two procedures and the number of false outliers identified in each sample by the two procedures as the dependent variables. The independent variables were the outlier identification procedure, the extremeness of the outliers, and the number of dimensions in which the outliers occurred. The data were entered in a repeated measures factorial design with procedure crossed with extremeness of outliers and with number of dimensions in which the outliers occurred. The data were analyzed according to the recommendations of Looney and Stanley (1989) and Barcikowski and Robey (1984).

The research design layout is shown is Table 2.

Table 2

Research Design Matrix

Dimensions	One				Two				Three			
Extremeness	3σ		6σ		3σ		6σ		3σ		6σ	
Procedure	1	2	1	2	1	2	1	2	1	2	1	2
$Sample_1$												
.												
.												
.												
$Sample_{1,100}$												

Tukey's Honestly Significant Difference (HSD) Procedure (SAS Institute, 1990c) was run on the main effects of procedure, dimensionality, and extremeness in order to compute the minimum significant differences.

RESULTS

The analysis of variance procedure was run on the data generated by the two outlier identification procedures to partition the variance accounted for by the design. As expected, the large sample size (n = 1,100) resulted in all the statistical tests being significant at the $p < 0.0001$ level; therefore, interpretation of the study is based on practical significance as indexed by eta-square rather than on statistical significance. In terms of practical significance only the main effects and one two-way interaction were found to be significant. Results were analyzed separately for the two dependent variables, the number of false outliers identified and the total number of outliers identified.

False Outliers

For the false outliers identified, the overall model and each of the variables were statistically significant at the 0.0001 level. The F-value for each variable is listed in Table 3. The overall model

accounted for 84.4% of the total variability; subtracting the variability accounted for by the sample (7.9%) leaves 76.2% accounted for by dimensionality (18.6%), degree of extremeness (47.3%), and the interaction of dimensionality and extremeness (10.3%). The procedure for identifying outliers accounted for less than 1% of the variability, as did the interactions of procedure with dimensionality, procedure with extremeness, and the three-way interaction of procedure, dimensionality, and extremeness.

Identifying false outliers represents error on the part of the outlier identification procedures. In a practical sense, the choice of procedure was not an issue; although statistically significant, the choice of procedure accounted for only about one-tenth of one percent of the variability. The degree of extremeness of the outliers accounted for most of the variability in the model. The six standard deviations and the three standard deviations degree of extremeness were significantly different at the 0.0001 level and accounted for about 47% of the variability. The dimensionality accounted for a significant amount of the variability. Tukey's HSD at the .05 level showed a minimum significant difference of 0.0271 between the two procedures and between the two degrees of extremeness and a minimum significant difference of 0.0397 among all three dimensions. Values obtained for Tukey's HSD can be found in Table 4.

Table 3

ANOVA Summary Table for False Outliers

Source	DF	Sum of Squares	Mean Square	F Value	R-Square
Model	1110	41215.99	37.13	58.85*	0.8438
Error	12089	7627.18	0.63		
Corrected Total	13199	48843.16			

Source	DF	Sum of Squares	Mean Square	F Value	R-Square
Dimensionality	2	9068.94	4534.47	7189.09*	0.1857
Extremeness	1	23108.05	23108.05	3984.48*	0.4731
Ext X Dim	2	5027.77	2513.89	3964.48*	0.1029
Procedure	1	54.61	54.81	86.55*	0.0011
Dim X Proc	2	32.50	16.25	25.75*	0.0007
Ext X Proc	1	34.93	34.93	55.36*	0.0007
Ext X Dim X Proc	2	26.28	13.14	20.83*	0.0005
Sample	1099	3862.91	3.51	5.57*	0.0791

* $p < 0.0001$

Table 4

Tukey's HSD for False Outliers

	Mean *	
Procedure	2 1.64727	Minimum significant difference
	1 1.51864	= 0.0271
Dimensionality	1 2.58409	Minimum significant
	2 1.61045	difference = 0.0397
	3 0.55432	
Extremeness	3 2.90606	Minimum significant
	6 0.25985	difference = 0.0271

* All means are significantly different

An inspection of several of the data sets showed that both outlier identification procedures selected similar numbers of false outliers in each of the six combinations of dimensionality and extremeness and many times selected the same observations as outliers. Table 5 lists the means and standard deviations for the number of false outliers identified by the procedures.

Table 5

Means and Standard Deviations for False Outliers

Procedure	Dimension	Extremeness	Mean*	Standard Deviation
Mahalanobis	1	3σ	4.38	1.43
Andrews-Pregibon	1	3σ	4.87	1.52
Mahalanobis	2	3σ	2.94	1.19
Andrews-Pregibon	2	3σ	3.06	1.24
Mahalanobis	3	3σ	1.05	0.90
Andrews-Pregibon	3	3σ	1.13	0.95
Mahalanobis	1	6σ	0.52	0.70
Andrews-Pregibon	1	6σ	0.56	0.73
Mahalanobis	2	6σ	0.20	0.40
Andrews-Pregibon	2	6σ	0.24	0.43
Mahalanobis	3	6σ	0.02	0.14
Andrews-Pregibon	3	6σ	0.02	0.14

* All means are based on *n*'s of 1,100

Total Outliers

For the total number of outliers identified, the overall model and each of the variables were statistically significant at the 0.0001 level. The F-value for each variable is listed in Table 6. The overall model accounted for 61% of the total variability; subtracting the variability accounted for by

the sample (17.5%) leaves 42.7% accounted for by dimensionality (10.7%), degree of extremeness (18.1%), and the interaction of dimensionality and extremeness (13.9%). The procedure accounted for less than 1% of the variability, as did the interactions of procedure with dimensionality, procedure with extremeness, and the three-way interaction of procedure, dimensionality, and extremeness.

Table 6

ANOVA Summary Table for Total Outliers

Source	DF	Sum of Squares	Mean Square	F Value	R-Square
Model	1110	18703.42	16.85	17.16*	0.6118
Error	12089	11868.27	0.9817		
Corrected Total	13199	30571.69			

Source	DF	Sum of Squares	Mean Square	F Value	R-Square
Dimensionality	2	3262.78	1631.39	1661.73*	0.1067
Extremeness	1	5534.18	5534.18	5637.11*	0.1810
Ext X Dim	2	4250.82	2125.41	2164.94*	0.1390
Procedure	1	89.51	89.51	91.18*	0.0029
Dim X Proc	2	74.07	37.03	37.73*	0.0024
Ext X Proc	1	63.70	63.70	64.89*	0.0021
Ext X Dim X Proc	2	64.91	32.45	33.06*	0.0021
Sample	1099	5363.44	4.88	4.97*	0.1754

* $p < 0.0001$

In identifying total outliers, the choice of outlier identification procedures was not an issue; although statistically significant, the choice of procedure accounted for less than three-tenths of one percent of the total variability. The degree of extremeness of the outliers accounted for most of the variability in the model; and the dimensionality accounted for a significant amount. Tukey's HSD indicated that at the .05 level there was a minimum significant difference of 0.0338 between the two degrees of extremeness and between the two outlier identification procedures. With a minimum significant difference of 0.0495, the two-dimension group was significantly different from the one- and three-dimension groups; the one- and three-dimension groups were not significantly different. Values obtained for Tukey's HSD can be found in Table 7.

Table 7

Tukey's HSD for Total Outliers

	Mean *	
Procedure	A 2 8.98970	Minimum significant difference
	B 1 8.82500	= 0.0338
Dimensionality	A 2 9.61045	Minimum significant
	B 1 8.55727	difference = 0.0495
	B 3 8.55432	
Extremeness	A 3 9.55485	Minimum significant
	B 6 8.25985	difference = 0.0338

* Means with the same letter are not significantly different

An inspection of several of the data sets showed that both outlier identification procedures selected all the induced outliers in the six standard deviation group and in the two and three dimension groups of three standard deviations. The two procedures selected the same observations as

outliers in the one dimension three standard deviation group. The one dimension three standard deviation group is the only place in which the procedures did not find all the induced outliers. Table 8 lists the means and standard deviations for the total number of outliers identified by the procedures in the various combinations of dimensionality and extremeness.

Table 8

Means and Standard Deviations for Total Outliers

Procedure	Dimension	Extremeness	Mean*	Standard Deviation
Mahalanobis	1	3σ	8.22	2.28
Andrews-Pregibon	1	3σ	8.93	2.10
Mahalanobis	2	3σ	10.94	1.19
Andrews-Pregibon	2	3σ	11.06	1.24
Mahalanobis	3	3σ	9.05	0.90
Andrews-Pregibon	3	3σ	9.13	0.95
Mahalanobis	1	6σ	8.52	0.70
Andrews-Pregibon	1	6σ	8.56	0.73
Mahalanobis	2	6σ	8.20	0.40
Andrews-Pregibon	2	6σ	8.24	0.43
Mahalanobis	3	6σ	8.02	0.14
Andrews-Pregibon	3	6σ	8.02	0.14

* All means are based on *n*'s of 1,100

Conclusions and Discussion

Two multivariate outlier identification procedures, the Mahalanobis distance and the Andrews-Pregibon statistic, were compared under varying conditions of extremeness and dimensionality. Data were generated from a multivariate normal population with known means and variances. Outliers were induced in the 1,100 samples of data according to six possible combinations of extremeness, three or six times the standard deviation of the population, and dimensionality, outliers occurring in one, two, or three dimensions. The two outlier identification procedures were run on each of the samples. Analysis of variance was run once using the number of false outliers identified as the dependent variable and again using the number of total outliers identified as the dependent variable. The independent variables in each case were outlier identification procedure, extremeness, and dimensionality. In each case, every variable was statistically significant at the $p < 0.0001$ level due to the large number of samples; therefore, analyzing the results based on practical significance as indexed by eta-square was determined to be more appropriate. Tukey's Honestly Significant Difference was calculated for each of the main effects.

In a practical sense, the choice of outlier identification procedure did not make a difference in regards to either false outliers or total outliers. In each case the procedure accounted for less than one percent of the total variability in the model.

The degree of extremeness of the outliers accounted for most of the variability in each case, with the three standard deviation group accounting for more outliers than the six standard deviation group.

The dimensionality of the outliers also accounted for a significant amount of the total variability. All three groups were significantly different in the false outlier model; the two-dimension group was significantly different from the other two groups in the total outlier model.

All of the interactions were statistically significant, but they accounted for a very small proportion of the variance.

Conclusions

With the two procedures in this study, the Mahalanobis distance and the Andrews-Pregibon statistic, there was a significant difference at the 0.0001 level in detecting both total outliers and false outliers; however, statistical significance was anticipated due to the large number of samples. In terms of practical significance, the outlier identification procedure accounted for less than one percent of the variability in the model; therefore, other than ease of calculation, there is no reason to choose one procedure over the other procedure. Some statistical packages, such as SPSSX, calculate the Mahalanobis distance, while no commercial package calculates the Andrews-Pregibon statistic. However, using SAS/IML (SAS Institute, 1990b), both procedures can be calculated quickly and efficiently.

The extremeness of the outliers, three or six standard deviations, accounted for 47.3% of the variability using false outliers as the dependent variable and 18.1% of the variability using total outliers as the dependent variable. Both degrees of extremeness were found to be significant at the 0.0001 level; there was a significant difference between the two degrees of extremeness in the number of outliers detected. With either false outliers or total outliers as the dependent variable, the three standard deviation degree of extremeness showed an honestly significant difference from the six standard deviation degree of extremeness. The mean number of outliers identified in the three standard deviation group was higher than that in the six standard deviation group for both false and total outliers.

The dimensionality, outliers occurring in one, two, or three dimensions, accounted for 18.6% of the variability using false outliers as the dependent variable and 10.7% of the variability using total outliers as the dependent variable. All three of the dimensions were found to be significantly

different at the 0.0001 level. There was a significant difference among the one-, two-, and three-dimension groups in the number of false outliers detected. There was an honestly significant difference between the two-dimension group and the one- and the three-dimension groups in the number of total outliers detected.

Discussion

It is important to realize that an "outlier" as identified by these two procedures is simply an observation that does not fit the distribution of the other scores. It must be verified that the observations which are identified are true outliers based on an error of measurement, recording, coding, or a deviation from the design.

The choice of an outlier detection procedure does not seem to be a major consideration for the researcher in a situation similar to what was modeled here. It is important for the researcher to be familiar with the data being analyzed and with the procedure being used. He or she must be able to take the observations that are identified as possible outliers by the selected procedure and to examine those observations in order to determine their validity. The decision of whether to accommodate or to reject an observation identified as an outlier must be a careful one. Both procedures in this study identified valid data points as outliers. The presence of these false outliers underscores the necessity of examining the outliers on an individual basis in order to verify that the observation is, in fact, an outlier. Deleting or weighting an observation strictly on the basis of an outlier identification procedure could lead to a loss of valid data. Accepting the results of a procedure without further investigation could lead to invalid results and inferences.

Both procedures identified data points three standard deviations from the mean as outliers. Although the probability of an observation three standard deviations from the mean occurring in normally distributed data is small, the observation could still be from the same distribution; in fact, not having any observations three standard deviations from the mean could signal a problem with the

data. Three and six standard deviations were chosen as the degrees of extremeness for this study. In retrospect, three standard deviations was probably not the best choice.

Recommendations for Further Research

Other procedures should be studied as they might produce different results. Many procedures are available, including several that are graphical. The decision of which procedure to use should be based on the researcher's knowledge of the procedure, the computer resources available and the type of analysis to be done. Several procedures are tailored to specific types of analyses, such as regression; therefore, if the researcher has a certain analysis in mind, he or she should investigate the procedures available in that area.

Studies should be done using data involving more dimensions. Procedures are fairly straightforward when there are only two dimensions, but the results become more obscured as the number of dimensions increases.

Studies should be done using different degrees of extremeness. A point six standard deviations from the mean is an outlier, but a point only three standard deviations from the mean is probably a valid data point.

An operational definition of an outlier is needed to facilitate research in this area. In the literature there is no consensus definition of an outlier, much less an operational definition. If the definition must be specified for each study done, the number of available procedures will continue to mount and comparison of results will be difficult. If a common operational definition could be developed, researchers could concentrate on the development of more adequate procedures, possibly through the refinement of existing procedures.

REFERENCES

Andrews, D. F. & Pregibon, D. (1978). Finding the outliers that matter. Journal of the Royal Statistical Society B, 1, 85-93.

Barcikowski, R. S. & Robey, R. R. (1984). Decisions in single group repeated measures analysis: Statistical tests and three computer packages. The American Statistician, 38, 148-150.

Barnett, V. (1978). The study of outliers: Purpose and model. Applied Statistics, 3, 242-250.

Barnett, V., & Lewis, L. (1978). Outliers in statistical data. Chichester: John Wiley & Sons.

Chatterjee, S. & Hadi, A. S. (1988). Sensitivity analysis in linear regression. New York: John Wiley & Sons.

Comrey, A. L. (1985). A method for removing outliers to improve factor analytic results. Multivariate Behavioral Research, 20, 273-281.

Dixon, W. J. (1950). Analysis of extreme values. Annals of Mathematical Statistics, 21, 488-506.

Douzenis, C. & Rakow, E. A. (1987). Outliers: A Potential Data Problem. Memphis, TN: Memphis State University. (ERIC Document Reproduction Service No. ED 291 798)

Elashoff, J. D. & Elashoff, R. M. (1970). A model for quadratic outliers in linear regression. Stanford University, CA: Stanford Center for Research and Development in Teaching. (ERIC Document Reproduction Service No. ED 047 020)

Gentleman, J. F. & Wilk, M. B. (1975). Detecting outliers. II. Supplementing the direct analysis of residuals. Biometrics, 31, 387-410.

Gnanadesikan, R. (1977). Methods for statistical data analysis of multivariate observations. New York: John Wiley & Sons.

Gnanadesikan, R. & Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics, 28, 81-124.

Gray, J. B. (1989). On the use of regression diagnostics. The Statistician, 38, 97-105.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. Technometrics, 11, 1-21.

Guttman, I. (1973). Care and handling of univariate or multivariate outliers in detecting spurosity - A Bayesian approach. Technometrics, 15, 723-738.

Hawkins, D. M. (1980). Identification of outliers. New York: Chapman and Hall.

Huber, P, J. (1977). Robust statistical procedures. Philadelphia: Society for Industrial and Applied Mathematics.

International Business Machines. (1988a). VM/XA System Product Interpreter: Reference Manual, First Edition, March 1988: Author.

International Business Machines. (1988b). VM/XA System Product Interpreter: User's Guide, First Edition, March 1988: Author.

Jarrell, M. G. (1991). Generating an empirical probability distribution for the Andrews-Pregibon statistic. Paper presented at the annual meeting of the Mid-South Educational Research Association, November, 1991, Lexington, KY.

Krzanowski, W. J. (1988). Principles of multivariate analysis. Oxford: Clarendon Press.

Looney, S. W. & Stanley, W. B. (1989). Exploratory repeated measures analysis for two or more groups. The American Statistician, 43(4), 220-225.

Morris, J. D. (1975). A computer program to create a population with any desired centroid and covariance matrix. Educational and Psychological Measurement, 35, 707-710.

Pascale, P. J. & Lovas, C. M. (1976). A computer program for detection of statistical outliers. Educational and Psychological Measurement, 36, 209-211.

Portnoy, S. (1988). Regression quantile diagnostics for multiple outliers. Unpublished manuscript. University of Illinois.

Rasmussen, J. L. (1988). Evaluating outlier identification tests: Mahalanobis D squared and Comrey Dk. Multivariate Behavioral Research, 23, 189-202.

Robertson, C. (1987). The Detection of outliers in free recall lists: An exploratory analysis. British Journal of Mathematical and Statistical Psychology, 40(2), 140-156.

Rousseeuw, P. J. & Leroy, A. M. (1987). Robust regression and outlier detection. New York: John Wiley & Sons.

Rousseeuw, P. J. & van Zomeren, B. C. (1990). Rejoinder to Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association, 85, 648-651.

SAS Institute Inc. (1990a). SAS Procedures Guide, Version 6, Third Edition. Cary, N.C.: Author.

SAS Institute Inc. (1990b). SAS/IML Software: Usage and Reference, Version 6, First Edition. Cary, N.C.: Author.

SAS Institute Inc. (1990c). SAS/STAT User's Guide: Volume 2, GLM-VARCOMP, Version 6, Third Edition. Cary, N.C.: Author.

Schwager, S. J. & Margolin, B. H. (1982). Detection of multivariate outliers. The Annals of Statistics, 10, 943-954.

Seber, G. A. F. (1984). Multivariate observations. New York: John Wiley & Sons.

Shapiro, S. S., Wilk, M. B., & Chen, M. J. (1968). A comparative study of various tests for normality. Journal of the American Statistical Association, 63, 1343-1372.

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. Psychological Bulletin, 95, 334-344.

Tabachnick, B. G. & Fidell, L. S. (1983). Using Multivariate Statistics. New York: Harper & Row, Publishers.

Wainer, H. (1976). Robust statistics: A survey and some prescriptions. Journal of Educational Statistics, 1, 285-312.

Wood, F. S. (1983). Measurements of observations far-out in influence and/or factor space. Paper presented at the Econometrics and Statistics Colloquium at the Chicago Graduate School of Business, Chicago, IL.